Influence of Sample Size in Land Cover Classification Accuracy Using Random Forest and Sentinel-2 Data in Portugal

Daniel Moraes [1 2], Pedro Benevides[2], Hugo Costa [12], Francisco D. Moreira[2], Mário Caetano [1 2]

[1] NOVA Information Management School (NOVA IMS), Universidade Nova Lisboa, Campus de Campolide, 1070 -312 Lisbon, Portugal

[2] Direção -Geral do Território, Rua da Artilharia Um, 107, 1099 -052 Lisboa, Portugal

This is the accepted version of the conference paper published by IEEE at *IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium: Proceedings:*

How to cite: Moraes, D., Benevides, P., Costa, H., Moreira, F. D., & Caetano, M. (2021). Influence of Sample Size in Land Cover Classification Accuracy Using Random Forest and Sentinel-2 Data in Portugal. In IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium: Proceedings (pp. 4232-4235). IEEE. https://doi.org/10.1109/IGARSS47720.2021.9553924

© 2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

INFLUENCE OF SAMPLE SIZE IN LAND COVER CLASSIFICATION ACCURACY USING RANDOM FOREST AND SENTINEL-2 DATA IN PORTUGAL

Daniel Moraes[1,2], Pedro Benevides[2], Hugo Costa[1,2], Francisco D. Moreira[2], Mário Caetano[1,2]

[1] NOVA Information Management School (NOVA IMS), Universidade Nova Lisboa, Campus de Campolide, 1070-312 Lisbon, Portugal
[2] Direção-Geral do Território, Rua da Artilharia Um, 107, 1099-052 Lisboa, Portugal

ABSTRACT

Classification accuracy of remote sensing images with supervised learning depends on the quality and characteristics of training samples. Size is a key aspect of a sample and its impact on classification depends on several factors, including the classifier employed, dimension on the feature space and land cover characteristics. Random Forest classifier is considered to be of low sensitivity to variations in sample size. However, further investigation is required when feature spaces are large and training is performed with spectral subclasses of the land cover classes to be mapped. This paper proposes to assess the impact of sample size in the classification accuracy of Random Forest using multi-temporal Sentinel-2 data and a detailed set of training subclasses to produce a map with general land cover classes. The results revealed similar classification accuracies after major reductions in sample size.

Index Terms— Random Forest, Sentinel-2, training sample, sample size

1. INTRODUCTION

Land cover and land use (LCLU) is considered an element of extreme relevance for the description and study of the environment [1]. LCLU and its derived products can benefit society in a range of areas, such as disasters, climate, water and agriculture [2]. Therefore, developing methods to map and quantify LCLU and its changes over time is essential. Remote sensing techniques have been widely adopted to map and monitor LCLU in a variety of spatial and temporal scales [3]. The European Space Agency (ESA) Sentinel-2 mission has global coverage, high spatial resolution (10, 20 and 60 m), 13 spectral bands and 5 day revisit time. The short revisit time results in a higher probability of acquiring cloud free images, thus supporting analysis of dense intra-annual time-series.

The recent advances in technology and broader data availability established a new LCLU mapping paradigm [4], including the emergence of automated data processing workflows using state-of-the-art machine learning algorithms to map LCLU over large areas [5]. Random Forest (RF) has received special interest within the remote sensing community, being successful to map LCLU with multi-dimensional data, simple in terms of parameters configuration [6] and not very sensitive to noise in the training data [7].

Supervised classification requires collecting training samples, and their characteristics can have a significant impact on classification accuracy. Regarding sample size, it is unclear how changes in the number of samples can affect accuracy. Overall, literature lacks advice on the minimum sample size, despite the broad understanding that increasing sample size results in higher accuracy [7]. Moreover, an adequate size may vary according to the classifier, number of predictor variables, trained classes, and size and spatial variability of the mapping region [8]. In terms of classifiers, RF was found to be significantly less sensitive to reductions in training sample size in comparison with single decision trees [9]. In addition, experiments conducted by [10] revealed that a reduction of 95% in the number of sampling units resulted in a decrease of less than 5% in accuracy. However, these studies were conducted on classifications with limited number of predictor variables.

Sampling for training image classification is traditionally a human dependent, costly and time consuming activity. Thus, automated processes based on existing reference datasets have been developed to extract training samples [11], which contributes to overcome some limitations of manual collection, for instance allowing the collection of a larger number of sampling units. The usefulness of automatic training extraction to produce large samples is, however, questionable when the classifier used has shown in the past to be effective with small samples sizes as is the case of Random Forest. Investigation is needed in the context of large multi-spectral and multi-temporal data, which increases the complexity of the feature space.

This paper adopts a strategy which considers two LCLU class nomenclatures, one used in the training process and the other corresponding to the final map nomenclature. The training classes are spectral subclasses of the map nomenclature, an approach employed to ensure that the spectral variability of the map nomenclature is taken into account at the training stage. Hence, this paper aims to conduct experiments to assess the influence of the size of the

training subclasses in classification accuracy, considering the map nomenclature. The classification is conducted with a Random Forest supervised classifier, using multi-temporal Sentinel-2 data and a semi-automatic workflow to extract training samples from existing reference datasets in a study region in Portugal.

2. DATA AND METHODS

This study was developed within the region of Trás-os-Montes, in the North of Portugal. The area comprises 11,778 km² and is characterized by mountainous areas occupied with rocks, forest and bushes, in addition to agriculture in the lower lands.

Sentinel-2 data from the agricultural year of 2018 (October 2017 to September 2018) were acquired to generate monthly composites and spectro-temporal metrics. Level-2A images were downloaded from the Theia Land Data Centre. In total, 457 images with less than 50% cloud cover were acquired. Pixels contaminated by clouds were converted to missing data and monthly composites were generated by calculating the median value of 10 bands (B2, B3, B4, B5, B6, B7, B8, B8A, B11 and B12), from which 5 spectral indices were computed. In addition, 7 spectro-temporal metrics were computed for each band and index. The final composite consisted of 285 bands: 10 bands and 5 indices for each month and 7 metrics for each band and index.

Reference data, namely the national land use and land cover map of Portugal for 2015 and 2018 (COS 2015, COS 2018), the Portuguese Land Parcel Identification System (LPIS) of 2018 and the OpenStreetMap (OSM) roads network of Portugal were used to delineate regions from which training samples were collected automatically. Filtering data were employed to refine the process. The Copernicus Land Monitoring Service's High Resolution Layers (HRL) products from 2015, national maps of burned areas and a mask of NDVI changes detected in 2015-2018 [12] were used as filtering datasets. A similar methodology was applied by [13]. The filters are employed in order to detect discrepancies between datasets and thus prevent mislabels in the training data. Some classes, however, needed to be trained manually as preliminary results indicated that some classes have low accuracies when sampled automatically. Manual training was based on manual delineation of polygons through visual interpretation of an ortophoto map of 2018 with 25 cm spatial resolution.

The automatic and manual training samples were extracted from the corresponding filtered or manual data sets, but subject to spatial constraints. A negative buffer of 40 m was applied to the automatic and manual data sets, and areas smaller than 1000 m² were deleted before sampling extraction. The reference datasets for automatic and manual training included a total of 22 LCLU classes. Table 1 presents the correspondence between classes used for RF training and the LCLU map nomenclature, the method of sample collection and the number of polygons that resulted from the filtering process.

Table 1: Correspondence between the LCLU map nomenclature and classes used in RF training, methods of training sample collection (A: automatic; M: manual) and number of training polygons.

LCLU map nomenclature	LCLU classes used in RF training	Method	Number of polygons
Built up	Built up	A	223
	Industrial	M	322
	Road network	A	-
Agriculture	Wheat	A	303
	Rye	A	751
	Oat	A	1146
	Ryegrass	A	34
	Triticale	A	66
	Corn	A	460
	Sunflower	A	1
	Barley	A	22
	Managed Grasslands	A	840
Natural Grasslands	Agricultural Natural Grassland	M	100
	Mountain Natural Grassland	M	47
Eucalyptus	Eucalyptus Adult	A	16
Other Broadleaf	Other Broadleaf	A	211
Maritime Pine	Maritime Pine	A	872
Other Coniferous	Other Coniferous	A	140
Shrubland	Dense Shrubland	M	255
Non-vegetated surfaces	Baresoil	A	453
	Bare Rock	M	953
Water	Water	A	492

The automatic and manual datasets were used to randomly extract training samples of varying size. Eight scenarios were tested in which 50, 500, 1000, 2000, 3000, 4000, 5000 and 6000 training sample units were extracted per class. In some scenarios there were few classes whose number of sample units was less than desired, and for those cases the entire number available was considered. Then, eight RF models corresponding to each sample size scenario were trained. The classification was implemented in Python, using the Scikit-learn library [14], parameterized with 500 trees and $\sqrt{n}$ as the number of features available at each node (n = 285).

The trained models were used to classify an independent validation dataset. This was composed of 535 sampling units drawn from stratified random sampling and manually labeled by visual interpretation of the ortophoto map already mentioned. The labels were assigned considering a 3x3 pixel window, with the sampling unit being located in the central pixel. This approach aims to address possible spatial displacement of the Sentinel-2 composite. For each validation sampling unit one or more reference class labels were allocated, when adequate (e.g. transition between two land cover patches). A sampling unit is considered correctly

classified if the class predicted by the RF classifier matches one of the labels assigned to that sample.

The accuracy assessment was conducted considering the 10 classes of the LCLU map nomenclature. The classifications were compared to evaluate whether variations in sample size affected classification accuracy. The accuracy estimators of [15] were used.

3. RESULTS AND DISCUSSION

The results of the eight classification scenarios exhibited fairly similar accuracies, despite the substantially different sample sizes (Fig. 1). The highest accuracy (73.7%) was achieved with 2000 sampling units per class, whereas the lowest accuracy (71.5%) was observed using 6000 sampling units per class. The variation in accuracy was ~2% and it is not possible to identify a trend in accuracy as a function of the size of the training. The uncertainty of the accuracy estimates is ~5% and the confidence intervals overlap, meaning that the differences between the classifications' accuracy are statistically insignificant.

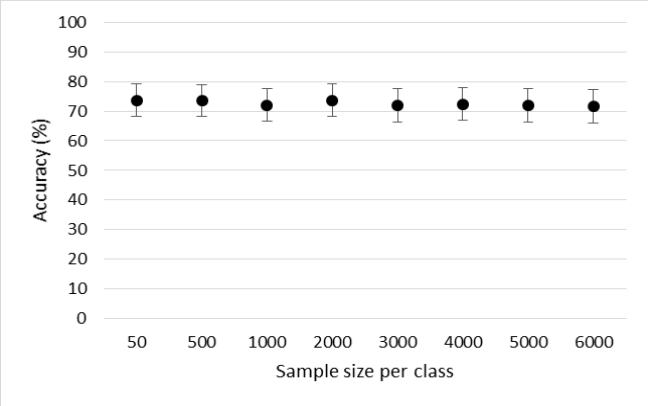

Fig. 1: Classification accuracy vs sample size.

These results are in accordance with the findings of [11, 12], which concluded that RF has low sensitivity to reduction in sample size. The outcomes indicate that smaller samples might be as capable as larger samples to properly discriminate land cover classes. As the majority of the classes have a large number of training polygons, sample units are collected from scattered areas potentially representative of various spectral conditions. Furthermore, the use of training subclasses ensure that spectral diversity is included in the samples regardless of their size. The histogram of the coefficient of variation (CV) computed for all bands and training classes is compared in Fig. 2. It illustrates that samples of 50 and 6000 units per class have similar CV distribution. A closer examination of the CVs of the near-infrared band for three distinct months considering the classes other broadleaf (OB), maritime pine (MP), other coniferous (OC), agricultural natural grassland (ANG), mountain natural grassland (MNG) and dense shrubland (DSB) (Table 2) also shows similar values when comparing distinct sample sizes.

The data also reveals that variability is similar regardless of the training polygons being generated automatically or manually. Figure 3 exhibits classification maps, revealing a fair similarity between the classifications with 50 and 6000 sampling units per class.

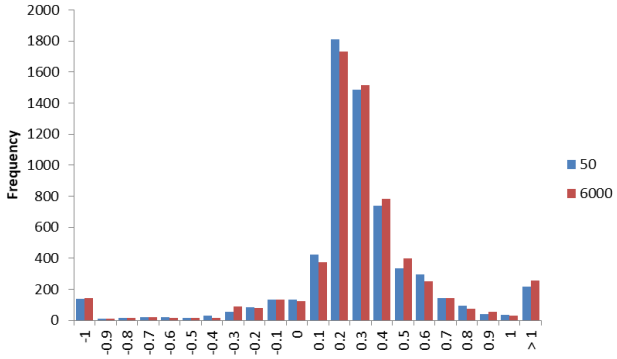

Fig. 2: Distribution of CVs computed for all bands and classes for samples of 50 and 6000 units per class.

Table 2: Comparison of CVs for near-infrared band per month for samples of 50 and 6000 units per class.

Class	Oct		Feb		Jul	
	50	6000	50	6000	50	6000
OB	0.22	0.23	0.20	0.22	0.10	0.10
MP	0.18	0.16	0.20	0.18	0.14	0.12
OC	0.13	0.16	0.14	0.17	0.13	0.17
ANG	0.22	0.20	0.17	0.19	0.15	0.16
MNG	0.13	0.14	0.13	0.17	0.10	0.12
DSB	0.18	0.17	0.26	0.23	0.17	0.16

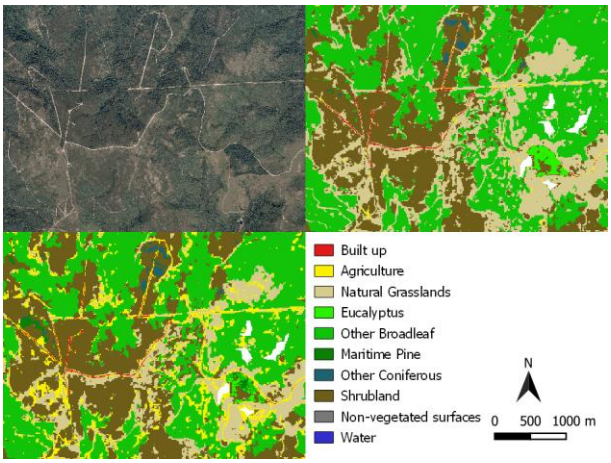

Fig. 3: Ortophoto map and classifications with 50 (top right) and 6000 (bottom left) sampling units per class.

4. CONCLUSION

This work conducted experiments with supervised multidimensional classification using Random Forest and multi-temporal Sentinel-2 data embedded in a semi-

automatic process of training samples collection to evaluate the impact of sample size in classification accuracy.

The results, in convergence with previous studies conducted with the same classifier, revealed that classification accuracy was fairly similar, even after a reduction of over 90% in the number of training sample units. Spectral variability analysis of the smaller and larger samples pointed to comparable values between both samples, suggesting that the smaller sample was as capable as the larger sample to discriminate land cover classes. This may be explained by the training strategy based on spectral subclasses, which ensures spectral diversity in the samples for all sizes. Additionally, the comparison of the CVs between automatically and manually collected samples for the near-infrared band indicated that there is no substantial difference in spectral variability for small and large samples.

The experiments ratified RF's low sensitivity to variations in sample size, illustrating that an increase in sample size does not necessarily yields higher classification accuracy for classifications with complex feature space and training spectral subclasses. This may have practical implications on the design of operational mapping workflows that currently tend to implement automatic processes in general, including automatic training sample collection. Collecting large samples automatically may seem advantageous, but, in the case of Random Forest, it may not afford higher classification accuracy when training subclasses ensure spectral diversity. However, it is worth mentioning that collecting smaller training samples might result in missing some subclasses, i.e. collecting a non-statistically complete representation of some thematic classes, which could affect the results.

4. ACKNOWLEDGEMENTS

The work has been supported by project foRESTER (PCIF/SSI/0102/2017), SCAPEFIRE (PCIF/MOS/0046/ 2017) and by Centro de Investigação em Gestão de Informação (MagIC), all funded by the Portuguese Foundation for Science and Technology (FCT). Value-added data processed by CNES for the Theia data centre www.theia-land.fr using Copernicus products. The processing uses algorithms developed by Theia's Scientific Expertise Centres.

5. REFERENCES

[1] M. Herold, J.S. Latham, A. Di Gregorio and C.C. Schmullius, "Evolving standards in land cover characterization", *Journal of Land Use Science*, vol. 1, no. 2-4, pp. 157-168, 2006.

[2] M.A. Wulder, J.C. White, S.N. Goward, J.G. Masek, J.R. Irons, M. Herold, W.B. Cohen, T.R. Loveland and C.E. Woodcock, "Landsat continuity: Issues and opportunities for land cover monitoring", *Remote Sensing of Environment*, vol. 112, no. 3, pp. 955-969, 2008.

[3] J. Cihlar, "Land cover mapping of large areas from satellites: Status and research priorities", *International Journal of Remote Sensing*, vol. 21, no. 6-7, pp. 1093-1114, 2000.

[4] M.A. Wulder, N.C. Coops, D.P. Roy, J.C. White and T. Hermosilla, "Land cover 2.0", *International Journal of Remote Sensing*, vol. 39, no. 12, pp. 4254-4284, 2018.

[5] T. Hermosilla, M.A. Wulder, J.C. White, N.C. Coops and G.W. Hobart, "Disturbance-Informed Annual Land Cover Classification Maps of Canada's Forested Ecosystems for a 29-Year Landsat Time Series", *Canadian Journal of Remote Sensing*, vol. 44, no. 1, pp. 67-87, 2018.

[6] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions", *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24-31, 2016.

[7] A.E. Maxwell, T.A. Warner and F. Fang, "Implementation of machine-learning classification in remote sensing: an applied review", *International Journal of Remote Sensing*, vol. 39, no. 9, pp. 2784-2817, 2018.

[8] C. Huang, L.S. Davis and J.R.G. Townshend, "An Assessment of Support Vector Machines for Land Cover Classification", *International Journal of Remote Sensing*, vol. 23, no. 4, pp. 725-749, 2002.

[9] V.F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo and J.P. Rigon-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification*", ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 67, pp. 93-104, 2012.

[10] P. Thanh Noi and M. Kappas, "Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery", *Sensors*, vol. 18, no. 1, pp. 18, 2018.

[11] J. Inglada, A. Vincent, M. Arias, B. Tardy, D. Morin and I. Rodes, "Operational high resolution land cover map production at the country scale using satellite image time series", *Remote Sensing*, vol. 9, no. 1, pp. 95, 2017.

[12] C. Hugo, P. Benevides, F. Marcelino, M. Caetano, "Introducing automatic satellite image processing into land cover mapping by photo-interpretation of airborne data", *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, pp. 29-34, 2020.

[13] I. Hernandez, P. Benevides, H. Costa and M. Caetano, "Exploring Sentinel-2 for land cover and crop mapping in Portugal", *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 43, pp. 83-89, 2020.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: machine learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[15] S.V. Stehman, "Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes", *International Journal of Remote Sensing*, vol. 35, no. 13, 2014.